Semi-Supervised Learning Approach for Indonesian Named Entity Recognition (NER) Using Co-Training Algorithm

Bayu Aryoyudanta
Electrical and Computer Engineering
Gadjah Mada University
Yogyakarta
Email: me@yudanta.web.id

Teguh Bharata Adji
Electrical and Computer Engineering
Gadjah Mada University
Yogyakarta
Email: adji@ugm.ac.id

Indriana Hidayah
Electrical and Computer Engineering
Gadjah Mada University
Yogyakarta
Email: indriana.h@ugm.ac.id

Abstract—The problem of utilizing machine learning approachin Indonesian Named Entity Recognition (NER) system is the limited amount of labelled data for training process. However, unlike the limited availability of labelled data, unlabelled data is widely available from many sources. This enables a semi- supervised learning approach to solve this NER system problem. This research aims to design a semi-supervised learning modelto solve NER system problem. A semi-supervised co-training learning is used to utilize unlabelled data in NER learning process to produce new labelled data that can be applied to enhance a new NER classi□cation system.This research uses two kinds of data, Indonesian DBPedia data as labelled data and news article text from Indonesian news sites (kompas.com, cnnindonesia.com, tempo.co, merdeka.com and viva.co.id) as unlabelled data. The pre-processing steps applied to analyze unstructured text are sentence segmentation, tokenization, stemming, and PoS Tagging. The results of this pre- process are the NER and its context used as unlabelled data for the semi-supervised co-training process. The SVM algorithm is used as a classi□cation algorithm in this process. 10 Cross Fold Validation is used as the system testing approach. Based on the result of the NER testing system, the precision is 73.6%, the recall is 80.1% and f1 mean is 76.5%.

***Index Terms*—NER, Bahasa Indonesia, Labelled Data, Unlabelled Data, Semi-Supervised Learning, Co-Training, SVM, Precision, Recall, F1Score**

I. Introduction

A named entity recognition (NER) is one sub task of Information Extraction (IE) domains in Natural Language Processing (NLP). NER system aims to identify and classify an entity based on its context. Common entity classification in a NER system consist of Organization, Person, Location, Date, Time, Currency, Percentage, Facility and Geo Political Entities (GPE)[1].

Nowadays Indonesian NER systems commonly use a traditional model applying a rule-based approach and machine learning model using supervised learning approach. One of NER systems using a rule-based approach is developed by Budi[2]. This system gives an accurate result in classifying the entity, but it has poor result in identifying the number of entities that is found. Unlike the traditional approach, the machine learning approach supervised by learning model tends to have better results in Indonesian NER system. Wiwin et al [3] in their Indonesian Medical NER using SVM classification can achieve nearly 90% score in accuracy. Luthfi et al [4] used SVM classification and DBPedia data for the training data in their NER system and the precision and recall scores obtained from the evaluation of the system are above 90%.

The supervised learning model gives better results, but it has limited labelled data for the training process. This problem occurs because it is difficult to get the sufficient amount of labelled data. The characteristics of labelled data are "costly" and "effortful". In contrast, unlabelled data is widely available. Unlabelled data can be easily found such as online in news article. The wide availability of unlabelled data enables semi-supervised learning approach to utilize unlabelled data for enhancing NER system.

The multi-view co-training algorithm used to utilize the unlabelled data for the learning process was introduced by Blum and Mitchell in [5]. The multi-view co-training they had developed was able to classify 788 web pages with 95% of accuracy by only using 12 labelled data.

This research aims to develop a semi-supervised learning model to solve Indonesian NER system problem. The multi-view co-training concept is used in this research to utilize unlabelled data in new NER system. The new labelled data obtained from the use of semi-supervised learning is applied to develop the existing Indonesian NER system.

II. Literature Review

A. Named Entity Recognition

A Named Entity (NE) is a noun phrase referring to one individual type. The process of identifying and classifying entities in a text into their class classifications is the most important part of IE. This process is commonly called Named Entity Recognition(NER)[6]. NER classifies an entity that has proper noun label (NNP) and numeric label (CDO, CDC, CDP, CDI) according to its context[7]. Common categories of NER class classifications are organization, person, location, date, time, money, percent, facility, and GPE[1]. The process of NER can be seen in Fig. 1.

978-1-5090-1709-6/16/$31.00 ©2016 IEEE

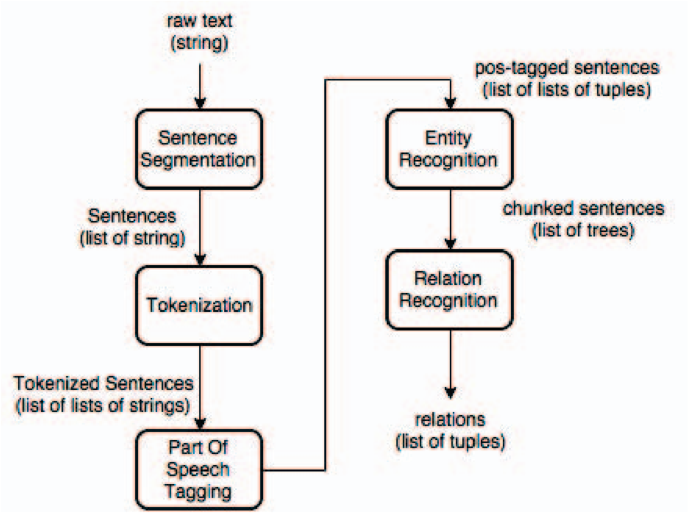

Fig. 1. Information Extraction Flow Diagram [1]

The NER system is preceded with text pre-processing process. The pre-processing steps applied to analyze the unstructured text are sentence segmentation, tokenization, stemming, and PoS Tagging.

1) Sentence Segmentation: The sentence segmentation processes raw-text into proper sentences using the rule of sentence. The output of this process is a list of sentences.

2) Tokenization: After the sentence segmentation process, the sentences are then split into smallest tokens in the form of words and punctuations. The example of an output tokens is: `['Pasangan', 'Tasdi-Tiwi', 'resmi', 'diusung', 'lima', 'partai', 'politik', ',', 'yakni', 'PDI', 'Perjuangan', ',', 'Partai', 'Gerindra', ',', 'Partai', 'Keadilan', 'Sejahtera', ',', 'Partai', 'Amanat', 'Nasional', 'dan', 'Partai', 'Nasional', 'Demokrat', '.']`.

3) Part of Speech Tagging: Part of Speech Tagging (PoS Tagging) is a process of labelling a part-of-speech tag or lexicon attribute into token in a text. PoS Tagging has an important task in NER system because this process leads the NER system to find the important entities with the proper noun (NNP) part-of-speech tag label. In this research, this process uses a HMM PoS Tagger developed by Wicaksono[8] that has been added with stemming mechanism to overcome the out of vocabulary problem.

Bigram HMM PoS Tagging used in this research refers to equations (1), (2), and (3)[8].

$$BigramHMM = \prod_{i=1}^{n} P(t_i|t_{i-1}).P(w_i|t_i) \tag{1}$$

Transition after smoothing

$$P(t_i|t_{i-1}) = \alpha P(t_1|t_{i-1}) + (1-\alpha).P(t_i) \tag{2}$$

Emission after smoothing

$$P(w_i|t_i) = \alpha P(w_i|t_i) + (1-\alpha).\frac{1}{n(w_i)} \tag{3}$$

The example of PoS Tagging process output is: `Pasangan/NN Tasdi-Tiwi/NN resmi/JJ diusung/VBT lima/CDP partai/NN politik/NN ,/, yakni/VBT PDI/NNP Perjuangan/NNP ,/, Partai/NNP Gerindra/NNP ,/, Partai/NNP Keadilan/NNP Sejahtera/NNP ,/, Partai/NNP Amanat/NNP Nasional/NNP dan/SC Partai/NN Nasional/JJ Demokrat/NNP ./.`

The obtained named entities from the PoS Tagging process are:

1) `PDI/NNP Perjuangan/NNP`
2) `Partai/NNP Gerindra/NNP`
3) `Partai/NNP Keadilan/NNP Sejahtera/NNP`
4) `Partai/NNP Amanat/NNP Nasional/NNP`

B. Multi-view Co-Training

The co-training algorithm was introduced in [5]. The idea of co-training is utilizing the unlabelled data to make a better clasification system with the use of limited amount of labelled data. Co-training uses two independent views as the learning process. The multi-view co-training assumes that:

1) Each view alone is sufficient to make a good classification, given enough labelled data as formulated is equation (4)[9].

$$x = [x^1, x^2] \tag{4}$$

2) The two views are *conditionally independent* given the class label, see equations (5) and (6)[9].

$$P(x^1|y.x^2) = P(x^1|y) \tag{5}$$

$$P(x^2|y.x^1) = P(x^2|y) \tag{6}$$

Co-training conducts the learning process for each view using the available labelled data. This learning process results in a prediction of the learned unlabelled data. A label is then given to the unlabelled data based on the most suitable prediction to make new labelled data. Next, this new labelled data is used to predict the other unlabelled data in the next iteration process[10]. The process of co-training algorithm can be found in Fig. 2.

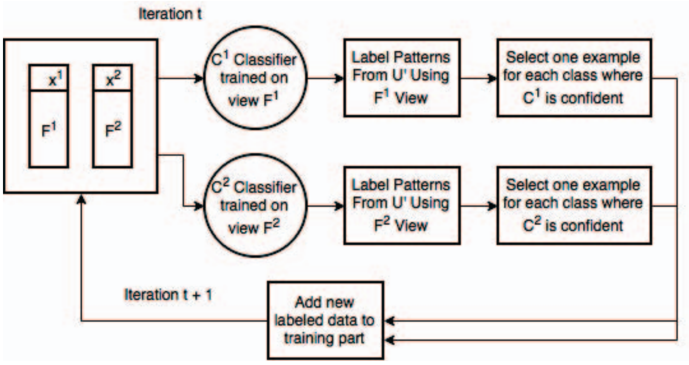
Iteration t
x¹
F¹
x²
F²
C¹ Classifier trained on view F¹
Label Patterns From U' Using F¹ View
Select one example for each class where C¹ is confident
C² Classifier trained on view F²
Label Patterns From U' Using F² View
Select one example for each class where C² is confident
Iteration t + 1
Add new labeled data to training part

Fig. 2. Co-Training Algorithm Work Flow [5]

Co-training then became an inspiration for related classification system research using semi-supervised model. Kiritchenko[11] developed an e-mail classification system using co-training approach, comparing the use of SVM and Bayesian as the classification algorithm. The result obtained using the SVM was better than that using Bayesian. Another co-training application for text classification was developed by Park Seong-Bae et al[12] and this research showed a better classification result of the use of co-training algorithm as well.

In the named entity classification process, an entity is classified according to its context. The examples are as follow.

Instance 1:

`bandara (Adi Sucipto) = FACILITY`
x^1`: Adi Sucipto, `x^2`: bandara dan y: FACILITY`

Instance 2:

`(Adi Sucipto) seorang pejuang = PERSON`
x^1`: Adi Sucipto, `x^2`: seorang pejuang dan y: PERSON`

In a formal NER system, each named entity is constructed by two views. A named entity is x^1 or view 1 and a context is x^2 or view 2. Each entity can be formulated as $x = [x^1, x^2]$, the same as the first assumption of co-training in equation (4). This condition is a "conditionally independent" because view 1 and view 2 are referring to each other.

The example of the use of multi-views assumption to solve the NER problem can be seen in Table I.

TABLE I
CO-TRAINING LABELLED AND UNLABELLED TRAINING INSTANCE SAMPLE

Instance	x^1	x^2	y
1	Adi Sucipto	Bandara	FACILITY
2	Bapak Adi Sucipto	Pejuang	PERSON
3	Soekarno-Hatta	Bandara	?
4	Bapak Purnomo	Teman Sejawat	?

1) From the labelled data in instance 1, it can be seen that the context "bandara" is classified as facility.
2) If the instance 1 is true, it can be concluded that instance 3 "Soekarno-hatta" is classified as facility, as we know that the context "bandara" is a facility.
3) The same assumption can be applied in instance 4 in the same way. The word "Bapak" is found in instance 2 and instance 4. Therefore, it can be concluded that the named entity "Bapak Purnomo" is classified as person.

A NER system has two views and each view is conditionally independent to each other. This state suits the requirements of co-training assumptions. Therefore, co-training algorithm is suitable in this research.

C. NER Evaluation

Information retrieval has adapted several standart of evaluation metrics, including precision, recall and combinet metric called F1Score. Precision score is a measure of how much information returned by the system is actually correct (equation (7)). Recall score is a measure of how much relevant information has extracted from the text (equation (8)).

$$P = \frac{|TP|}{|TP + FP|} \tag{7}$$

$$R = \frac{|TP|}{|TP + FN|} \tag{8}$$

$$F1Score = \frac{2.PR}{P + R} \tag{9}$$

The precision and recall score are antagonistic to each other. This situation has led to the use of combined measure called F1Score. The F1Score balances recall and precision and this combined metric calculation became the performance index of NER system (equation (9)) [13].

III. METHODOLOGY

A. Design System

The design system of this research has two parts. The first part is the co-training process. The purpose of this part is to collect new labelled data from unlabelled data. The second part is testing the new NER system. The new NER system utilizes new labelled data from the co-training process. In both co-training process and NER system, SVM algorithm is used in the classification process. The design system is described in Fig. 3.

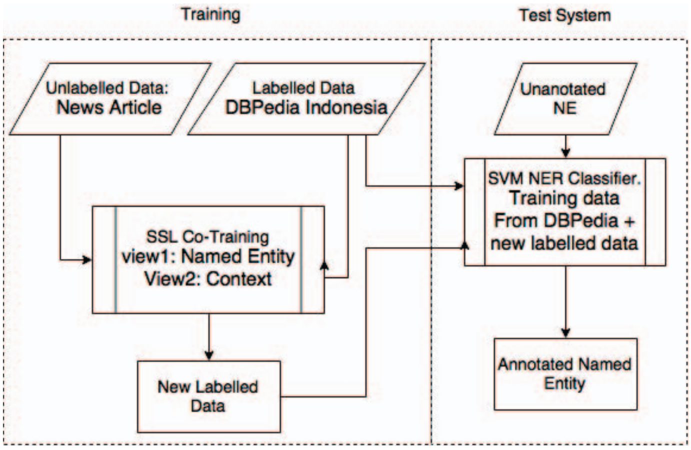
Training
Test System
Unlabelled Data: News Article
Labelled Data DBPedia Indonesia
Unanotated NE
SSL Co-Training view1: Named Entity View2: Context
SVM NER Classifier. Training data From DBPedia + new labelled data
New Labelled Data
Annotated Named Entity

Fig. 3. Proposed Semi-supervised Learning NER System

The co-training pseudo code used in the learning process can be seen in Fig. 4.

B. Labelled and Unlabelled Data

This research uses two kinds of data, labelled data used as the supervised data and unlabelled data used as the semi-supervised learning data. The labelled data is taken from Indonesian DBPedia articles that has named-entity and abstract information. The example of labelled data used as the supervised data can be found in Table II. The unlabelled data is taken from online news articles from cnnindonesia.com,

Fig. 4. Proposed Semi-supervised Algorithm

```
Algorithm 1: NER multi-views learning using co-training
  Labelled = $\{x_i, y_i\}_{i=1}^{l}$ $Unlabelled = \{x_j\}_{j=1}^{u}$;
  each x have two view $[x^1, x^2]$;
  initialize $L_1 = L_2 = \{x_i, y_i\}_{i=1}^{l}$ ;
  for $n := 1$ to $n := 5$ do  n iteration
      for do every unlabelled data $\{x_j\}_{j=1}^{u}$
          - try to classify x unlabelled data with $f(x)^1$ and
          $f(x)^2$ separately;
          - use $f(x)^1$ prediction to help $x^2$ prediction, and
          add these labelled data to $L_2$;
          - use $f(x)^2$ prediction to help $x^1$ prediction, and
          add these labelled data to $L_1$;
          - remove new labelled data from unlabelled list;
          - reload each classifier ($f(x)^1$ and $f(x)^2$) with
          new $L_1$ and $L_2$ data.
      end
  end
```

tempo.co, merdeka.com, viva.co.id, and kompas.com. The example of unlabelled data used as semi-supervised learning process can be found in Table III.

TABLE II
LABELLED DATA

Named Entity (view1)	Airlangga Sucipto
Sentence (view2)	Airlangga Sutjipto yang dipanggil “Angga” adalah seorang pemain sepak bola Indonesia. Ia berposisi sebagai penyerang. Saat ini ia memperkuat club Persib Bandung. Ia memiliki tinggi badan 167 sentimeter dan bermassa 67 kilogram. Ia pernah membela timnas Indonesia U-20 dan timnas Indonesia di SEA Games XXIV di Thailand. Ayahnya bernama Bambang Sutjipto dan Ibunya benama Yati Sumaryati. Mantan pemain Deltras Sidoarjo pada tahun 2008.
Tag	PERSON

TABLE III
UNLABELED DATA

Named Entity (view1)	Pakusadewo
Konteks (view2)	Tio - Julie
Sentence	Pemain : Tio Pakusadewo, Julie Estelle, Widyawati, Rio Dewanto, Chicco JerikhoProduser : Anggia Kharisma, Chicco Jerikho, Angga D.
Tag	?

C. Testing

This research uses 1468 data as the labelled data. The data is then classified as person (365), location (347), organization (326), facility (363), location (347), date (50) and GPE (17). The unlabelled data consists of 564 entities and divided into two data sets, 90% for training and 10% for testing.

The evaluation process is conducted in both training and testing systems to obtain the precision score, recall score and F1Score. 10 Cross Fold Validation approach is used in the testing process. This process produces average and maximum performance from NER system using semi-supervised learning approach.

IV. Results And Discussions

The result of the co-training and evaluation process can be seen in table IV and table V. From the evaluation of the Indonesian NER system developed using semi-supervised learning approach, the precision score is 0.736, the recall score is 0.801 and F1Score is 0.765. If those scores are converted into percentage, the precision is 73.6%, the recall is 80.1%, and F1Score is 76.5%.

TABLE IV
TRAINING RESULT

Fold	TP	FP	FN	Precision	Recall	F1Score
1	329	94	85	0.777	0.794	0.786
2	334	97	77	0.774	0.812	0.793
3	345	95	68	0.784	0.835	0.808
4	334	102	72	0.766	0.822	0.793
5	345	102	61	0.771	0.849	0.808
6	337	95	76	0.780	0.815	0.797
7	329	100	79	0.766	0.806	0.786
8	338	97	73	0.777	0.822	0.799
9	323	92	93	0.778	0.776	0.777
10	331	101	76	0.766	0.813	0.789
			AVG	0.774	0.814	0.793
			MAX	0.784	0.849	0.808

TABLE V
TESTING RESULT

Fold	TP	FP	FN	Precision	Recall	F1Score
1	36	14	6	0.720	0.857	0.782
2	36	15	5	0.705	0.878	0.782
3	34	11	11	0.755	0.755	0.755
4	38	14	4	0.730	0.904	0.808
5	44	7	5	0.862	0.897	0.880
6	30	10	16	0.750	0.652	0.697
7	29	16	11	0.644	0.725	0.682
8	32	15	9	0.680	0.780	0.727
9	30	14	12	0.681	0.714	0.697
10	41	8	7	0.836	0.854	0.845
			AVG	0.736	0.801	0.765
			MAX	0.862	0.904	0.880

A semi-supervised learning approach using co-training algorithm is compatible with the way NER system classifies entities, as it has 2 views of co-training assumptions, the named entity and the context, that are “conditionally independent” to each other. Therefore, semi-supervised approach enables this research to utilize unlabelled data in order to solve the problem of classifying entities using Indonesian NER system.

Based on the result of the NER system developed using co-training approach, it can be seen that the precision is 73.5%, the recall is 80.1%, and F1Score is 76.5%. Although the result of this research is not as high as that shown in the supervised model conducted by [3] and [4] that its evaluation score is almost 90%, the use of semi-supervised model enables the system to work with less labelled data and to form new labelled data using unlabelled data as the learning process. Moreover, semi-supervised learning allows the NER system to form more new labelled data through the learning process.

In a multi-view co-training process, view 1 and view 2 teach each other. The problem occurs if the context extraction for view 2 does not give a term or phrase that supports named entities that results in poor learning process.

A semi-supervised learning still has the possibility to provide incorrect label to new labelled data and it can be a problem in the further iteration because the system can generate more incorrect labelled data. Thus, it is necessary to add a correction mechanism for incorrect new labelled data produced by co-training.

V. Conclusion

A semi-supervised learning approach is suitable to be applied in developing an Indonesian NER system. Based on the result of the developed Indonesian NER system evaluation, the precision is 73.5%, the recall is 80.1%, and F1Score is 76.5%. Although the evaluation scores are not as high as those produced by NER system using supervised approach, Indonesian NER system using semi-supervised approach is capable of reducing "cost" and "effort" used in getting labelled data by utilizing unlabelled data as the learning process.

The problem of extracting the entities context used in the learning process that might be a challenge in the future research is developing an extraction context model from entities based on semantic ontology or graph. It is also considerably important to add a correction mechanism for new labelled data.

References

[1] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 2009. [Online]. Available: http://it-ebooks.info/book/261/

[2] I. Budi, "Association Rules Mining for Name Entity Recognition," *Information Systems Journal*, pp. 15–18, 2003.

[3] S. Wiwin, S. Iping, and P. Ayu, "imNER Indonesian Medical Named Entity Recognition," *2014 2nd International Conference on Technology, Informatics, Management, Engineering & Environtment Bandung, Indonesia*, no. 1, pp. 19–21, 2014.

[4] A. Luthfi, B. Distiawan, and R. Manurung, "Building an Indonesian named entity recognizer using Wikipedia and DBPedia," in *2014 International Conference on Asian Language Processing (IALP)*, 2014, pp. 19–22. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6973520

[5] T. Mitchell and A. Blum, "Combining labeled and unlabeled data with co-training," *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, 1998. [Online]. Available: http://dl.acm.org/citation.cfm?id=279943.279962

[6] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[7] H. Isozaki and H. Kazawa, "Efficient support vector classifiers for named entity recognition," *Proceedings of the 19th international conference on Computational linguistics*, pp. 1–7, 2002. [Online]. Available: http://dx.doi.org/10.3115/1072228.1072282

[8] a.F. Wicaksono and a. Purwarianti, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia," *Proceedings of the 4th International Malindo (Malaysia-Indonesia) Workshop*, no. August, 2010.

[9] X. Zhu and A. B. Goldberg, "Introduction to Semi-Supervised Learning," in *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2009, vol. 3, no. 1, pp. 1–130. [Online]. Available: http://www.morganclaypool.com/doi/abs/10.2200/S00196ED1V01Y200906AIM006

[10] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-Supervised Learning*. London: The MIT Press, 2006, vol. 1, no. 2. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21243728

[11] S. Kiritchenko and S. Matwin, "Email Classification with Co-Training," *Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative research*, p. 8, 2001. [Online]. Available: http://portal.acm.org/citation.cfm?id=782096.782104

[12] S. B. Park and B. T. Zhang, "Co-trained support vector machines for large scale unstructured document classification using unlabeled data and syntactic information," *Information Processing and Management*, vol. 40, no. 3, pp. 421–439, 2004.

[13] D. Jurafsky and J. H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition," *Speech and Language Processing An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition*, vol. 21, pp. 0–934, 2000. [Online]. Available: http://www.mitpressjournals.org/doi/pdf/10.1162/089120100750105975